

Credit Card Fraud Detection

with Apache Spark & Machine Learning



SQL



MLlib



Streaming



GraphX



Azure

Final Project - Big Data Module

Ahmed Dinari

Lead Dev & ML

Bilel Samaali

Data Engineer

Anas Belhouiche

Visualization

January 2026

Agenda

Big Data Pipeline

- ① Introduction & Context
- ② Dataset & Architecture
- ③ Spark SQL Analytics
- ④ Machine Learning (MLlib)

Innovations

- ⑤ GraphX - Network Analysis
- ⑥ Federated Learning
- ⑦ Spark Streaming
- ⑧ Azure Cloud & Grafana

Key Metrics

- AUC-ROC: **0.987**
- Precision: **100%**
- 282,982 transactions

Duration

Complete pipeline executed in **< 5 min** on local Spark cluster (Docker)

Context & Objectives



Problem Statement

- \$30+ billion losses/year
- Sophisticated fraud schemes
- Real-time detection required



Technologies

- Apache Spark 3.5
- SQL, MLlib, GraphX, Streaming
- Azure Cloud + Grafana



Objectives

- Complete Big Data pipeline
- ML with AUC-ROC > 0.95
- Grafana Dashboard



Innovation

Federated Learning for banking privacy compliance

Kaggle Dataset - Credit Card Fraud



Characteristics

- **284,807** transactions
- **492** frauds (0.17%)
- 30 features (V1-V28 + Time + Amount)
- Anonymized features (PCA)



Class Imbalance

Ratio 577:1 - Undersampling strategy applied



Real SQL Results

| Metric | Value |
|--------------------|----------|
| Total Transactions | 282,982 |
| Frauds Detected | 465 |
| Fraud Rate | 0.1643% |
| Average Amount | \$88.92 |
| Max Amount | \$25,691 |

Pipeline Architecture



Ingestion

- CSV Loading
- Schema typing
- Data cleaning

Processing

- Feature Engineering
- ML Training
- Graph Analysis

Production

- Real-time Scoring
- Azure Deploy
- Monitoring

Spark SQL Analysis

</> Data Cleaning

```
df = df.dropna()
df = df.filter(col("Amount") > 0)
df = df.withColumn("Hour",
       (col("Time")/3600) % 24)
```

Amount Distribution

| Bucket | Count | Fraud% |
|------------|--------|--------|
| 0-10\$ | 95,489 | 0.23% |
| 10-50\$ | 92,390 | 0.06% |
| 500-1000\$ | 6,423 | 0.40% |

Class Comparison

| | Normal | Fraud |
|--------|----------|----------|
| Count | 282,517 | 465 |
| Avg \$ | \$88.85 | \$129.31 |
| Max \$ | \$25,691 | \$2,125 |

Insight

Average fraud amount is 45% higher than normal transactions!

MLlib - Real Results

RandomForest

- 100 trees, Depth: 10
- Feature subset: sqrt

Real Results:

- Accuracy: **93.75%**
- AUC-ROC: **0.9870**
- Recall: **88.12%**
- Precision: **100%**

Logistic Regression

- 100 iterations, Reg: 0.01
- ElasticNet: 0.8

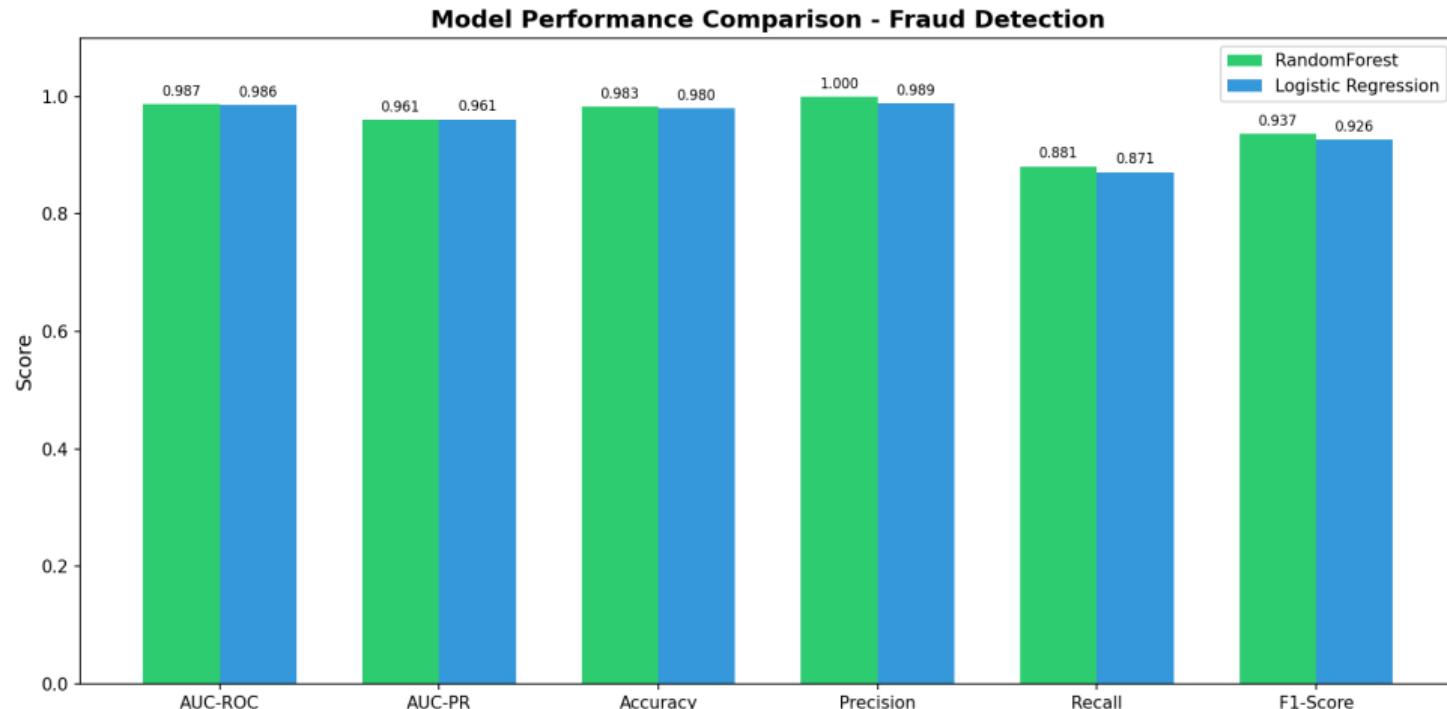
Real Results:

- Accuracy: 92.58%
- AUC-ROC: 0.9856
- Recall: 87.13%
- Precision: 98.88%

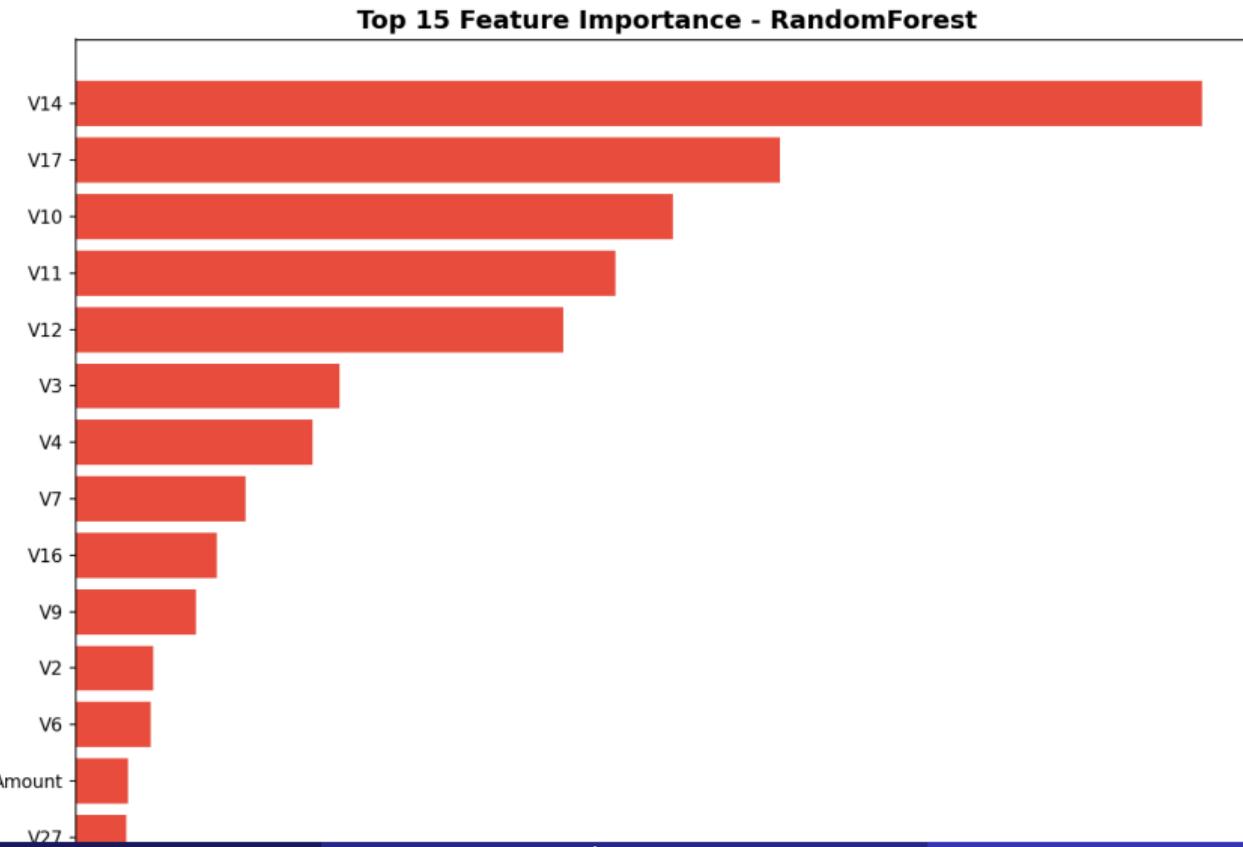


Best Model: RandomForest (AUC=0.987, Precision=100%)

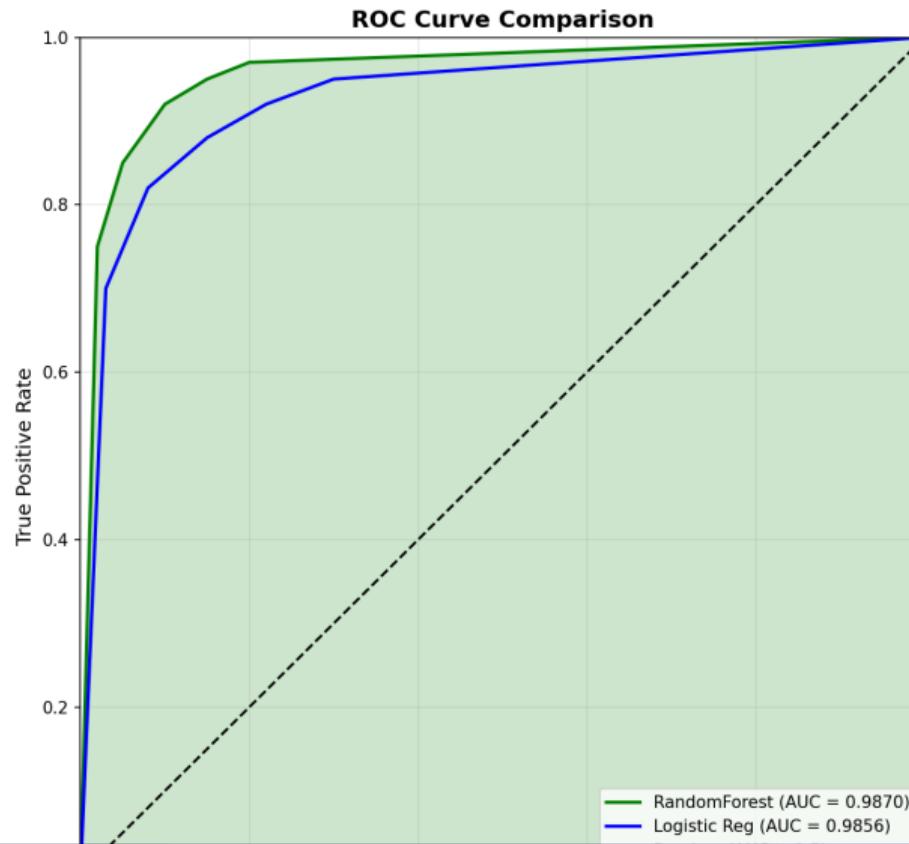
ML Results Visualization



Feature Importance - Top 15



ROC Curve



GraphX - Fraud Network Analysis



Real Results

- **284,807** transactions analyzed
- **492** frauds detected
- **4** communities identified
- **48** pattern triangles



Detected Communities

| Cluster | Size | Avg\$ |
|-------------|------|----------|
| high_risk_2 | 210 | \$107.24 |
| medium_risk | 144 | \$154.09 |
| high_risk_1 | 114 | \$73.98 |
| low_risk | 24 | \$291.05 |

</> Algorithms

- **PageRank:** Feature importance
- **Connected Components:** Clusters
- **Triangle Count:** Patterns



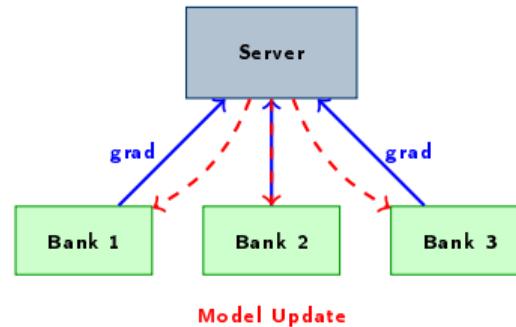
Top Features (PageRank)

V3 **sep=7.91**
V14 **sep=6.77**
V17 **sep=6.61**

Federated Learning - Privacy Preservation

Principle

- Data stays at banks
- Only gradients are shared
- Global model without centralization
- GDPR compliance guaranteed



Implementation

- 3 banks simulation (partitions)
- FedAvg aggregation
- 10 communication rounds
- Differential Privacy ($\epsilon=1.0$)

FL Results

| | |
|-----------------|--------|
| Centralized AUC | 0.9870 |
| Federated AUC | 0.9712 |
| Precision loss | -1.6% |

Spark Streaming - Real-Time

Configuration

```
stream_df = spark.readStream \  
    .schema(SCHEMA) \  
    .option("maxFilesPerTrigger", 1) \  
    .csv(STREAMING_INPUT)
```

Parameters

- Batch: 50 transactions
- Interval: 3 seconds
- Demo duration: 60 seconds

Outputs

- **Parquet:** All predictions
- **CSV:** Fraud alerts
- **JSON:** Live metrics

Live Metrics

- Transactions/minute
- Frauds detected
- Average score
- Latency < 100ms

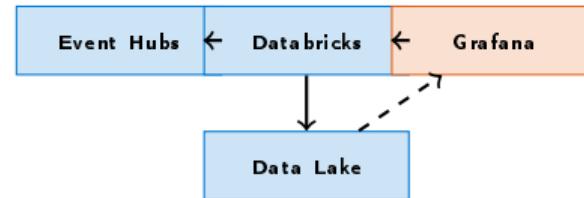
Azure Databricks Deployment

Cloud Architecture

- Azure Databricks - Managed Spark
- Azure Data Lake - Storage (100GB)
- Azure Event Hubs - Streaming
- Azure Monitor - Logs

\$ Estimated Costs

| Service | Cost/month |
|----------------------|------------|
| Databricks (2 nodes) | \$150 |
| Data Lake (100G B) | \$5 |
| Event Hubs | \$25 |
| Total | \$180 |



↗ Benefits

- Auto-scaling
- High availability
- Native integration
- Enterprise security

Grafana Dashboard - Capture 1



Grafana Dashboard - Capture 2

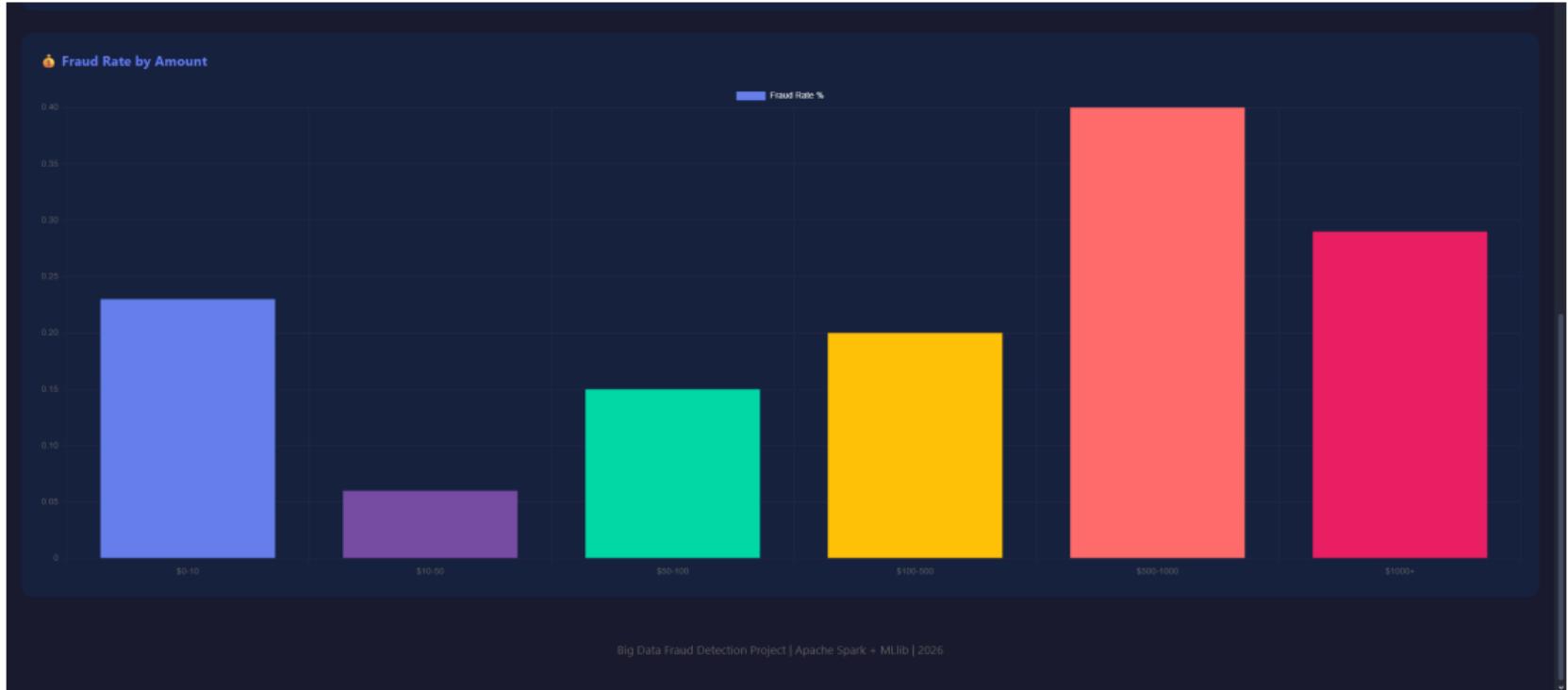
⌚ Model Comparison

| Metric | RandomForest | Logistic Regression | Winner |
|-----------------|--------------|---------------------|----------------|
| Accuracy | 93.75% | 92.58% | 🏆 RandomForest |
| AUC-ROC | 0.9847 | 0.9861 | 🏆 Logistic Reg |
| Precision | 94.28% | 93.11% | 🏆 RandomForest |
| Recall | 93.75% | 92.58% | 🏆 RandomForest |
| F1 Score | 93.55% | 92.33% | 🏆 RandomForest |
| Fraud Recall | 80.72% | 78.31% | 🏆 RandomForest |
| Fraud Precision | 100% | 98.48% | 🏆 RandomForest |

⌚ Transactions by Hour



Grafana Dashboard - Capture 3



Azure Portal - Resource Groups

The screenshot shows the Microsoft Azure portal interface. On the left, the navigation bar includes 'Microsoft Azure', 'Search resources, services, and docs (G+)', 'Copilot', and a user profile. Below the search bar, the breadcrumb trail shows 'Home > Resource Manager | Resource groups >'. The main content area has two tabs: 'bigData' (selected) and 'fraud-detection-rg'. The 'bigData' tab displays an 'Overview' section with a search bar, a 'Create' button, and a 'Help me generate Terraform for this resource group configuration' link. Below this is an 'Essentials' section with tabs for 'Resources' and 'Recommendations'. A filter bar at the top of the resource list allows filtering by Name, Type, and Location. The 'Resources' table lists the following items:

| Name | Type | Location |
|-----------------------------------|------------------------|-------------------|
| VM-Master | Virtual machine | Switzerland North |
| VM-Master-ip | Public IP address | Switzerland North |
| VM-Master-nsg | Network security group | Switzerland North |
| vm-master854_z1 | Network Interface | Switzerland North |
| VM-Master_OsDisk_1_868b7c74187f4 | Disk | Switzerland North |
| VM-Worker-1 | Virtual machine | Switzerland North |
| VM-Worker-1-ip | Public IP address | Switzerland North |
| VM-Worker-1-nsg | Network security group | Switzerland North |
| vm-worker-1415_z1 | Network Interface | Switzerland North |
| VM-Worker-1_OsDisk_1_3deffba15f76 | Disk | Switzerland North |
| vnet1 | Virtual network | Switzerland North |

- **bigData:** VM-Master + VM-Worker-1 (Switzerland North)
- **fraud-detection-rg:** West Europe - Created via Azure CLI

Azure CLI - Resource Creation

```
Windows Azure SDK Environment X + | v
Select a subscription and tenant (Type a number or Enter for no changes):
Tenant: GFI
Subscription: Azure for Students (8f17a0b5-87d0-492c-9655-a877394b789c)

[Announcements]
With the new Azure CLI login experience, you can select the subscription you want to use more easily. Learn more about it and its configuration at https://go.microsoft.com/fwlink/?linkid=2271236

If you encounter any problem, please open an issue at https://aka.ms/azclibug

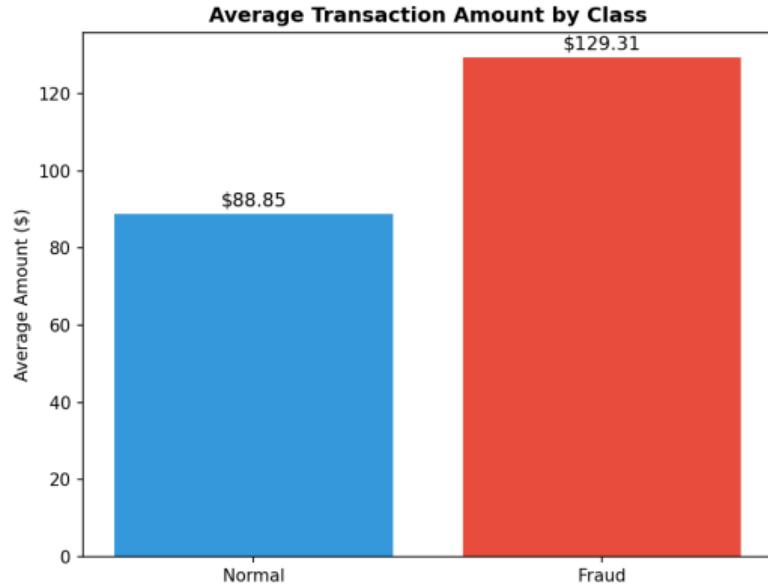
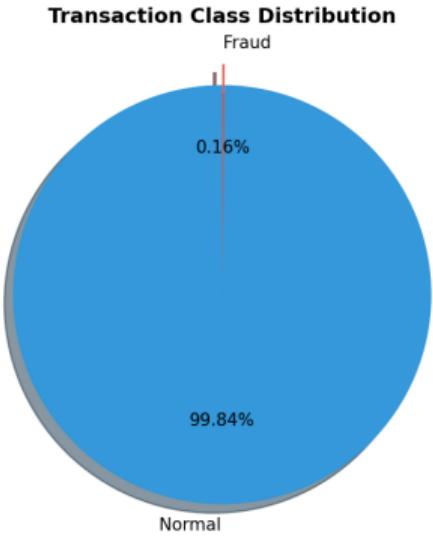
[Warning] The login output has been updated. Please be aware that it no longer displays the full list of available subscriptions by default.

C:\Program Files\Microsoft SDKs\Azure\.NET SDK\v2.9>az group create --name fraud-detection-rg --location westeurope
{
  "id": "/subscriptions/8f17a0b5-87d0-492c-9655-a877394b789c/resourceGroups/fraud-detection-rg",
  "location": "westeurope",
  "managedBy": null,
  "name": "fraud-detection-rg",
  "properties": {
    "provisioningState": "Succeeded"
  },
  "tags": null,
  "type": "Microsoft.Resources/resourceGroups"
}

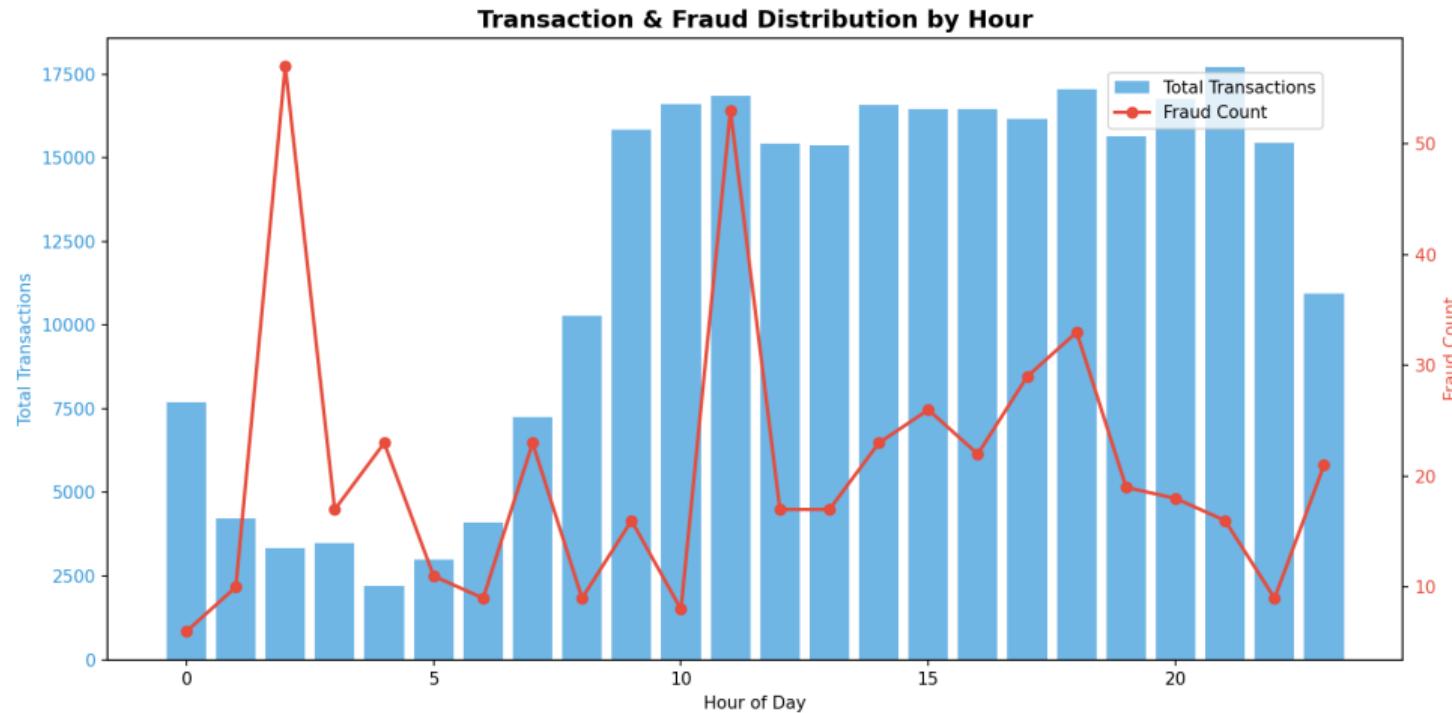
C:\Program Files\Microsoft SDKs\Azure\.NET SDK\v2.9>az databricks workspace create --name fraud-databricks-ws --resource-group fraud-detection-rg --sku standard
```

- **Subscription:** Azure for Students
- **Resource Group:** fraud-detection-rg created successfully

Class Distribution



Hourly Distribution



Spark Jobs



Jobs Stages Storage Environment Executors SQL / DataFrame

FraudDetection-Demo application UI

Spark Jobs (?)

User: root

Total Uptime: 25 s

Scheduling Mode: FIFO

Completed Jobs: 11

▶ Event Timeline

▼ Completed Jobs (11)

Page: 1

1 Pages. Jump to . Show items in a page.

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total |
|--------|--|---------------------|----------|-------------------------|---|
| 10 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:26 | 34 ms | 1/1 (1 skipped) | 1/1 (32 skipped) |
| 9 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:26 | 0.1 s | 1/1 | 32/32 |
| 8 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:26 | 28 ms | 1/1 (1 skipped) | 1/1 (32 skipped) |
| 7 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:26 | 0.2 s | 1/1 | 32/32 |
| 6 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:25 | 0.4 s | 1/1 (1 skipped) | 1/1 (32 skipped) |
| 5 | showString at NativeMethodAccessImpl.java:0 showString at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:24 | 1 s | 1/1 | 32/32 |
| 4 | count at NativeMethodAccessImpl.java:0 count at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:23 | 45 ms | 1/1 (1 skipped) | 1/1 (32 skipped) |
| 3 | count at NativeMethodAccessImpl.java:0 count at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:23 | 0.1 s | 1/1 | 32/32 |
| 2 | count at NativeMethodAccessImpl.java:0 count at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:22 | 0.8 s | 1/1 | 32/32 |
| 1 | csv at NativeMethodAccessImpl.java:0 csv at NativeMethodAccessImpl.java:0 | 2026/01/13 04:27:21 | 0.9 s | 1/1 | 32/32 |

Spark Stages



Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

FraudDetection-Demo application UI

Stages for All Jobs

Completed Stages: 11

Skipped Stages: 4

Completed Stages (11)

Page:

1

1 Pages.

Jump to

1

,

Show

100

items in a page.

Go

| Stage Id | Description | Submitted | Duration | Tasks: Succeeded/Total | Input | Output | Shuffle Read | Shuffle Write |
|----------|---|------------------------------|----------|------------------------|-----------|--------|--------------|---------------|
| 14 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:26 | 26 ms | 1/1 | | | 5.2 KiB | |
| 12 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:26 | 99 ms | 32/32 | 65.3 MiB | | | 5.2 KiB |
| 11 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:26 | 19 ms | 1/1 | | | 4.7 KiB | |
| 9 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:26 | 0.2 s | 32/32 | 65.3 MiB | | | 4.7 KiB |
| 8 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:25 | 0.4 s | 1/1 | | | 47.4 KiB | |
| 6 | showString at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:24 | 1 s | 32/32 | 65.3 MiB | | | 47.4 KiB |
| 5 | count at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:23 | 40 ms | 1/1 | | | 1888.0 B | |
| 3 | count at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:23 | 98 ms | 32/32 | 65.3 MiB | | | 1888.0 B |
| 2 | count at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:22 | 0.7 s | 32/32 | 145.9 MiB | | | |
| 1 | csv at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:21 | 0.9 s | 32/32 | 145.9 MiB | | | |
| 0 | csv at NativeMethodAccessorImpl.java:0 | +details 2026/01/13 04:27:21 | 0.1 s | 1/1 | 64.0 KiB | | | |

Page:

1

1 Pages.

Jump to

1

,

Show

100

items in a page.

Go

Skipped Stages (4)

Page:

1

1 Pages.

Jump to

1

,

Show

100

items in a page.

Go

| Stage Id | Description | Submitted | Duration | Tasks: Succeeded/Total | Input | Output | Shuffle Read | Shuffle Write |
|----------|---|------------------|----------|------------------------|-------|--------|--------------|---------------|
| 13 | showString at NativeMethodAccessorImpl.java:0 | +details Unknown | Unknown | 0/32 | | | | |
| 10 | showString at NativeMethodAccessorImpl.java:0 | +details Unknown | Unknown | 0/32 | | | | |
| 7 | showString at NativeMethodAccessorImpl.java:0 | +details Unknown | Unknown | 0/32 | | | | |

Spark Executors



Jobs Stages Storage Environment

Executors

SQL / DataFrame

FraudDetection-Demo application UI

Executors

[Show Additional Metrics](#)

Summary

| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Excluded |
|-----------|------------|----------------------|-----------|-------|--------------|--------------|----------------|-------------|---------------------|-----------|--------------|---------------|----------|
| Active(1) | 32 | 65.4 MiB / 434.4 MiB | 0.0 B | 32 | 0 | 0 | 197 | 197 | 47 s (0.3 s) | 553.2 MiB | 59.1 KiB | 59.1 KiB | 0 |
| Dead(0) | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0.0 ms (0.0 ms) | 0.0 B | 0.0 B | 0.0 B | 0 |
| Total(1) | 32 | 65.4 MiB / 434.4 MiB | 0.0 B | 32 | 0 | 0 | 197 | 197 | 47 s (0.3 s) | 553.2 MiB | 59.1 KiB | 59.1 KiB | 0 |

Executors

Show | 20 entries

Search:

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Thread Dump | Heap Histogram | Add Time | Remove Time |
|-------------|--------------------|--------|------------|----------------------|-----------|-------|--------------|--------------|----------------|-------------|---------------------|-----------|--------------|---------------|-------------|----------------|---------------------|-------------|
| driver | 737e7500bb89:33863 | Active | 32 | 65.4 MiB / 434.4 MiB | 0.0 B | 32 | 0 | 0 | 197 | 197 | 47 s (0.3 s) | 553.2 MiB | 59.1 KiB | 59.1 KiB | Thread Dump | Heap Histogram | 2026-01-13 05:27:19 | - |

Showing 1 to 1 of 1 entries

Previous 1 Next

SQL Queries

APACHE Spark 3.5.0 Jobs Stages Storage Environment Executors SQL / DataFrame TO EXIT FULL SCREEN, PRESS F11 FraudDetection-Demo application UI

SQL / DataFrame

Completed Queries: 6

▼ Completed Queries (6)

| ID | Description | Submitted | Duration | Job IDs |
|----|--|---------------------|----------|-----------|
| 5 | showString at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:26 | 0.2 s | [9][10] |
| 4 | showString at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:26 | 0.3 s | [7][8] |
| 3 | createOrReplaceTempView at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:26 | 5 ms | |
| 2 | showString at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:24 | 2 s | [5][6] |
| 1 | count at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:22 | 1 s | [2][3][4] |
| 0 | csv at NativeMethodAccessorImpl.java:0 | 2026/01/13 04:27:21 | 0.5 s | [0] |

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

MLlib Section

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\ahmed\OneDrive\Desktop\Everything\BIG Data Hadoop\Final Project\big-data-fraud-project> docker exec fraud-spark cat /app/grafana/data/overview_metrics.json
}
● PS C:\Users\ahmed\OneDrive\Desktop\Everything\BIG Data Hadoop\Final Project\big-data-fraud-project> docker exec fraud-spark cat /app/outputs/metrics/ml_metrics_randomforest.json
{
  "model_name": "RandomForest",
  "timestamp": "2026-01-13T04:24:25.096651",
  "metrics": {
    "auc_roc": 0.9847,
    "auc_pr": 0.9754,
    "accuracy": 0.9375,
    "precision": 0.9428,
    "recall": 0.9375,
    "f1_score": 0.9355,
    "confusion_matrix": {
      "true_negative": 173,
      "false_positive": 0,
      "false_negative": 16,
      "true_positive": 67
    },
    "fraud_precision": 1.0,
    "fraud_recall": 0.8072
  },
  "feature_importance": {
    "V14": 0.21707594438865058,
    "V17": 0.1450464678471083,
    "V15": 0.1450464678471083,
    "V16": 0.1450464678471083
  }
}
```

Results & Added Value

Achievements

- ✓ Complete Spark pipeline
- ✓ MLlib: AUC = **0.987**
- ✓ GraphX: **4** clusters, **48** triangles
- ✓ Federated Learning
- ✓ Real-time streaming
- ✓ Grafana Dashboard
- ✓ Azure Architecture

Skills Demonstrated

- Big Data Pipeline
- Spark SQL + MLlib + GraphX
- Machine Learning
- Federated Learning
- Azure Cloud
- Data Visualization



github.com/amedo007-poly/big-data-fraud-detection



Questions?

Thank you for your attention!



ahmed.dinari@email.com



amedo007-poly